

Sparse Estimation of Cox Proportional Hazards Models via Approximated Information Criteria

Xiaogang Su,^{1*} Chalani S. Wijayasinghe,¹ Juanjuan Fan,² and Ying Zhang^{3,4}

¹Department of Mathematical Sciences, University of Texas, El Paso, Texas, U.S.A.

²Department of Mathematics and Statistics, San Diego State University, San Diego, California, U.S.A.

³Department of Biostatistics, Indiana University Fairbanks School of Public Health and School of Medicine, Indianapolis, Indiana, U.S.A.

⁴Department of Statistics, Shanghai Jiao Tong University, Shanghai, China.

*email: xsu@utep.edu

SUMMARY. We propose a new sparse estimation method for Cox (1972) proportional hazards models by optimizing an approximated information criterion. The main idea involves approximation of the ℓ_0 norm with a continuous or smooth unit dent function. The proposed method bridges the best subset selection and regularization by borrowing strength from both. It mimics the best subset selection using a penalized likelihood approach yet with no need of a tuning parameter. We further reformulate the problem with a reparameterization step so that it reduces to one unconstrained nonconvex yet smooth programming problem, which can be solved efficiently as in computing the maximum partial likelihood estimator (MPLE). Furthermore, the reparameterization tactic yields an additional advantage in terms of circumventing postselection inference. The oracle property of the proposed method is established. Both simulated experiments and empirical examples are provided for assessment and illustration.

KEY WORDS: AIC; BIC; Cox proportional hazards model; Regularization; Sparse estimation; Variable selection.

1. Introduction

Consider the usual setup for censored survival data. Let (T'_i, C'_i) denote the failure and censoring times for the i th individual for $i = 1, \dots, n$. The observed failure time is $T_i = \min(T'_i, C'_i)$ with failure status indicated by $\delta_i = I\{T'_i \leq C'_i\}$. Let $\mathbf{z}_i \in \mathbb{R}^p$ denote the p -dimensional covariate vector associated with subject i . Without loss of generality (WLOG), we assume that all the covariates have been standardized. For identifiability concern in the ensuing modeling and inference, we assume that T'_i and C'_i are independent given \mathbf{z}_i , i.e., $T'_i \perp C'_i | \mathbf{z}_i$. Thus, the observed data consist of $\{(T_i, \delta_i, \mathbf{z}_i) : i = 1, \dots, n\}$. The Cox (1972) proportional hazards (PH) model formulates the hazard function of T'_i given \mathbf{z}_i as

$$h(t|\mathbf{z}_i) = h_0(t) \exp(\boldsymbol{\beta}^T \mathbf{z}_i), \quad (1)$$

where $\boldsymbol{\beta} = (\beta_j) \in \mathbb{R}^p$ is the unknown regression parameter vector. Estimation of model (1) is based on the partial likelihood (Cox, 1975). Throughout the article, we shall restrict our discussion to the traditional finite dimension scenario, i.e., p is fixed and $p < n$, while possible high-dimensional extensions will be discussed later. Assuming no or few ties in the observed failure times, the partial log-likelihood function for $\boldsymbol{\beta}$ is given by

$$l(\boldsymbol{\beta}) = \sum_{i=1}^n \delta_i \left[\mathbf{z}_i^T \boldsymbol{\beta} - \log \sum_{i'=1}^n \{I(T_{i'} \geq T_i) \exp(\mathbf{z}_{i'}^T \boldsymbol{\beta})\} \right].$$

Concerning variable selection, the true $\boldsymbol{\beta}$ is often sparse in the sense that some of its components are zeros. By “sparse

estimation,” we refer to methods and procedures that allow for identification of zero components in $\boldsymbol{\beta}$ and estimation of its nonzero components simultaneously.

There are two major types of variable selection techniques for survival models. Both can be generally formulated as a penalized partial likelihood form:

$$\min_{\boldsymbol{\beta}} -2l(\boldsymbol{\beta}) + \lambda \cdot \text{pen}(\boldsymbol{\beta}), \quad (2)$$

where $\text{pen}(\boldsymbol{\beta})$ is a penalty function and the penalty parameter λ is either fixed a priori or treated as a tuning parameter. The first type is the best subset selection (BSS) method, where a model selection criterion such as AIC (Akaike, 1974) or BIC (Schwarz, 1978) is employed to compare models of all choices. BSS solves

$$\min_{\boldsymbol{\beta}} -2l(\boldsymbol{\beta}) + \lambda_0 \|\boldsymbol{\beta}\|_0, \quad (3)$$

where the ℓ_0 norm $\|\boldsymbol{\beta}\|_0 = \text{card}(\boldsymbol{\beta}) = \sum_{j=1}^p I\{\beta_j \neq 0\}$ measures the model complexity and the penalty parameter λ_0 is fixed as $\lambda_0 = \ln(n_0)$ for BIC, with n_0 being the total number of uncensored failures (Vollinsky and Raftery, 2000). If AIC is used, then $\lambda_0 = 2$. Due to the discrete nature of the ℓ_0 norm, solving (3) is NP-hard and its optimization is proceeded in two steps: fit every model with the maximum partial likelihood method and then compare the fitted models according to an information criterion. Although faster algorithms such as branch-and-bound (Furnival and Wilson, 1974) and iterative hard thresholding (Blumensath and Davies, 2009) and

heuristic surrogates such as stepwise procedures are available for this combinatorial optimization problem, the best subset selection is infeasible for moderately large p .

The second type is regularization, as exemplified by Least Absolute Shrinkage and Selection Operator (LASSO; Tibshirani, 1997), adaptive LASSO (ALASSO; Zhang and Lu, 2007), and Smoothly Clipped Absolute Deviation penalty (SCAD; Fan and Li, 2002). LASSO replaces the ℓ_0 norm with the ℓ_1 norm for convex relaxation, i.e., $\text{pen}(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1 = \sum_{j=1}^p |\beta_j|$, so that the problem becomes

$$\min_{\boldsymbol{\beta}} -2l(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1. \quad (4)$$

The significance of LASSO is that it reformulates sparse estimation into a continuous convex programming problem. Nevertheless, the performance of LASSO is unsatisfactory in either variable selection or parameter estimation. To improve, ALASSO applies a weighted ℓ_1 norm and SCAD employs a nonconvex penalty, both enjoying the oracle property, i.e., consistency in selecting variables and efficiency in estimating the nonzero coefficients, under certain conditions given that λ can be appropriately chosen.

With the regularization approach, the penalty function does not correspond well to the model complexity $\|\boldsymbol{\beta}\|_0$. As a result, the value of the penalty parameter λ is no longer trackable by referring to AIC or BIC. Therefore, optimization of (4) has to resort to two steps as well: first solve (4) for every fixed $\lambda > 0$ to obtain a regularization path $\{\tilde{\boldsymbol{\beta}}(\lambda) : \lambda > 0\}$, and then select the best λ via an information criterion along the path. While two fast algorithms, homotopy (Osborne, Presnell, and Turlach, 2000) and coordinate descent (Fu, 1998), have been proposed for ℓ_1 regularization, the two-step procedure can be time-consuming when solving (4) with a fixed λ entails iterative procedures, which is the case in Cox PH models. Compared to BSS, the ℓ_1 regularization procedure amounts to seeking minimum AIC or BIC only along the regularization path, which is a much reduced search space since $\tilde{\boldsymbol{\beta}}(\lambda)$ is nothing but a one-dimensional curve indexed by λ in the original search space \mathbb{R}^p . Therefore, it is reasonable to deduce that the regularization-based estimators may not perform as well as the estimator obtained with BSS, if AIC or BIC is used as the performance criterion.

Besides the concerns about performance and computational efficiency, another major challenge that both methods face is postselection inference. Statistical inference is routinely done based on the nonzero coefficient estimates in the regularization or these selected variables in BSS. In such a practice, it has been taken for granted that model selection has no or little effect on the subsequent inferences, a myth recently debunked by Leeb and Pötscher (2005) who discussed an impossibility result for some postselection estimation. The problem can be manifested by the fact that no standard error results are available for zero coefficient estimates in regularization approaches. One is referred to Berk et al. (2013) and Lockhart et al. (2014) for further discussions and recent developments on this issue.

In this article, we put forward a new method of conducting sparse estimation for Cox PH models that helps address the aforementioned deficiencies. The main idea is to approxi-

mate the information criteria so that it yields a continuous or smooth objective function for easier optimization. For simplicity, we abbreviate the proposed method as MIC for “minimum information criterion.” MIC extends the best subset selection to scenarios with large p . At the same time, MIC can be regarded as a regularization method, yet free of tuning. In order to circumvent the postselection inference, we also propose a technical maneuver to obtain a valid statistical testing for parameters with zero estimates. The remainder of the article is organized as follows. Section 2 presents the proposed method in detail, as well as its asymptotic properties. Section 3 addresses the postselection inference problem. Section 4 contains numerical results based on both simulated experiments and a real data example. Section 5 ends the article with a short discussion.

2. Minimizing the Approximated Information Criterion

We seek a new sparse estimation method that can borrow strength from both BSS and regularization and bridge them. We start with BSS by approximating the discrete ℓ_0 norm and make further improvement by capitalizing on knowledge of regularization.

2.1. Approximation of ℓ_0 Norm

While the idea of optimization plays a critical role in both BSS and regularization, the primary motivation of our approach comes from approximation. The discrete nature of ℓ_0 norm in (3) poses the main obstacle for BSS, which motivates us to seek a continuous or smooth approximation to it with a continuous surrogate function. This essentially involves approximation of $I(\boldsymbol{\beta} \neq 0)$. For this purpose, we introduce the concept of unit dent functions. We call a continuous function $w : \mathbb{R} \rightarrow [0, 1]$ a *unit dent function at 0* if it satisfies: (i) $w(\cdot)$ is an even function such that $w(\beta) = w(-\beta)$; (ii) $w(0) = 0$ and $\lim_{|\beta| \rightarrow \infty} w(\beta) = 1$; and (iii) $w(\beta)$ is increasing on \mathbb{R}_+ . Denote by \mathcal{D}_0 the space of all unit dent functions at 0. It can be seen that \mathcal{D}_0 is closed under operations such as composition and product. Clearly, any unit dent function in \mathcal{D}_0 can be viewed as a continuous approximation of $I(\boldsymbol{\beta} \neq 0)$.

Among many others, one natural choice in \mathcal{D}_0 is the hyperbolic tangent function given by

$$\tanh(a|\beta|^r) = \frac{\exp(2a|\beta|^r) - 1}{\exp(2a|\beta|^r) + 1},$$

where $a > 0$ is a scale parameter that controls the sharpness of the approximation and $r \in \mathbb{N}$ has typical values of 1 and 2. Figure 1 plots the tanh function with $r = 1$ in (a) and $r = 2$ in (b), for different choices of $a = 1, 2, \dots, 200$. It can be seen that a relatively large a is needed in order to provide a good approximation. It is also interesting to note that the curve with $r = 2$ is smooth while the curve with $r = 1$ has a cusp at $\beta = 0$.

From the perspective of regularization, Fan and Li (2001) spelled out three desired properties for the penalty function: unbiasedness, sparsity, and continuity. It can be seen that both unbiasedness and continuity are easily satisfied by both choices. To enforce sparsity, the choice of $r = 1$ is favorable as opposed to $r = 2$ at the first sight. However, setting $r = 1$

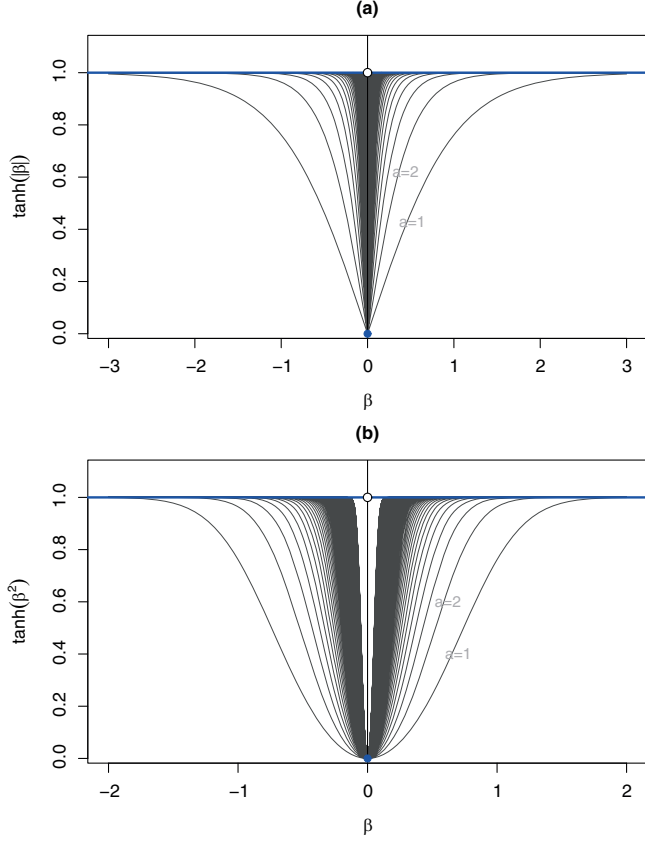


Figure 1. Hyperbolic tangent penalty functions $\tanh(a|\gamma|^r)$ that approximate the indicator function $I(\gamma \neq 0)$: (a) $r = 1$ and (b) $r = 2$. The value of a ranges from 1 to 200.

leads to a nonsmooth optimization problem and the resultant estimates suffer from the same postselection inference. These concerns motivate us to restrict our attention to $r = 2$ in order to ensure smoothness. We then devise a different way of enforcing sparsity and circumventing postselection inference.

2.2. Reparameterization

With $r = 2$, solving $\min_{\beta} \{-2l(\beta) + \lambda_0 \sum_{j=1}^p \tanh(a\beta_j^2)\}$ facilitates a surrogate BSS method as in (3). This is a smooth optimization problem; however, it does not provide sparse estimates. To remedy, we shall reparameterize the problem by introducing $\gamma = (\gamma_j) \in \mathbb{R}^p$, which relates to β as follows. Define $w_j = w(\gamma_j) = \tanh(a\gamma_j^2)$ for $j = 1, \dots, p$ and matrix $\mathbf{W} = \text{diag}(w_j)$. Then, set $\beta_j = \gamma_j w_j$ for each j . That is, we reparameterize β in terms of γ such that $\beta = \mathbf{W}\gamma$. Now, we consider the following optimization problem

$$\min_{\gamma} -2l(\mathbf{W}\gamma) + \lambda_0 \text{tr}(\mathbf{W}), \quad (5)$$

where $\text{tr}(\mathbf{W}) = \sum_{j=1}^p w_j$ is the trace of matrix \mathbf{W} . It turns out that this simple reparameterization step not only helps enforce sparsity while keeping the optimization problem smooth but also addresses the inference issue with zero estimates.

One original motivation of the above reparameterization step came from nonnegative garotte (NG; Breiman, 1995). NG is formulated as a sign-constrained regularization problem based on the decomposition $\beta = \text{sgn}(\beta) \cdot |\beta|$. Assuming that the signs of β can be correctly specified by another consistent estimator, say, the MPLE $\hat{\beta}$, it remains to estimate $|\beta|$. Reparameterizing $\beta = \text{diag}(\hat{\beta})\gamma$ for $\gamma = (\gamma_j)$ with $\gamma_j \geq 0$, NG first estimates γ by solving

$$\min_{\gamma} -2l(\beta) \quad \text{s.t.} \quad \sum_{j=1}^p \gamma_j \leq \tau \text{ and } \gamma_j \geq 0,$$

with τ being a tuning parameter, and then obtains the estimated regression coefficients $\hat{\beta}_j = \hat{\gamma}_j \hat{\beta}_j$ for $j = 1, \dots, p$. One fundamental problem with NG is that if any sign of the initial estimator $\hat{\beta}$ is wrongly specified, which occurs often with real data owing to multicollinearity or other complexities, then it becomes hopeless for NG to make correction. Comparatively, MIC is based on a different decomposition $\beta = \beta \cdot I\{\beta \neq 0\}$. Setting $\gamma = \beta$ and approximating $I\{\gamma \neq 0\}$ by $w(\gamma)$ lead to the reparameterization $\beta = \gamma w(\gamma)$, which does not depend on an initial estimate.

With the reparameterization, the regression coefficient vector remains β . However, the decision vector in (5) becomes γ . This helps keep the optimization problem smooth. We first obtain the estimate of γ , $\tilde{\gamma}$, and hence $\tilde{\mathbf{W}} = \text{diag}(w(\tilde{\gamma}_j))$, then we compute the estimate of β as $\tilde{\beta} = \tilde{\mathbf{W}}\tilde{\gamma}$. The function $w(\gamma)$ is smooth in γ with the first two derivatives given by $\dot{w} = dw/d\gamma = 2a\gamma(1 - w^2)$ and $\ddot{w} = 2a(1 - w^2)(1 - 4a\gamma^2 w)$. To see why (5) leads to sparse estimation of β , it is helpful to examine the penalty $w(\gamma) = \tanh(a\gamma^2)$ as a function of β . First of all, there is one-to-one correspondence between β and γ , as shown in Figure 2a. As a function of β , $w(\gamma)$ is a unit dent function that can be used to approximate its ℓ_0 norm. Applying the chain rule and differentiation of the inverse function yields

$$\frac{dw(\gamma)}{d\beta} = \frac{dw(\gamma)}{d\gamma} \cdot \left(\frac{d\beta}{d\gamma}\right)^{-1} = \frac{\dot{w}}{w + \gamma\dot{w}}.$$

Similar arguments can be applied to obtain its higher order derivatives. A closer look reveals that $w(\gamma)$ is a smooth function in β everywhere except at $\beta = 0$ where $d\beta/d\gamma = 0$. Figure 2b plots $w(\gamma)$ versus β , showing that $w(\gamma)$ possesses all the properties of the desired penalty function for sparse estimation of β .

2.3. Asymptotic Results

In this section, the asymptotic properties of the MIC estimator $\hat{\beta}$ are studied. Owing to the use of the counting processes and martingale theories, all the arguments hold for time-dependent covariates $\mathbf{z} = \mathbf{z}(t)$. WLOG, we work on the time interval $t \in [0, 1]$. Our notations follow those similar to Anderson and Gill (1982), Fan and Li (2002), and Zhang and Lu (2007). We consider the MIC estimator $\hat{\beta}$ as the solution

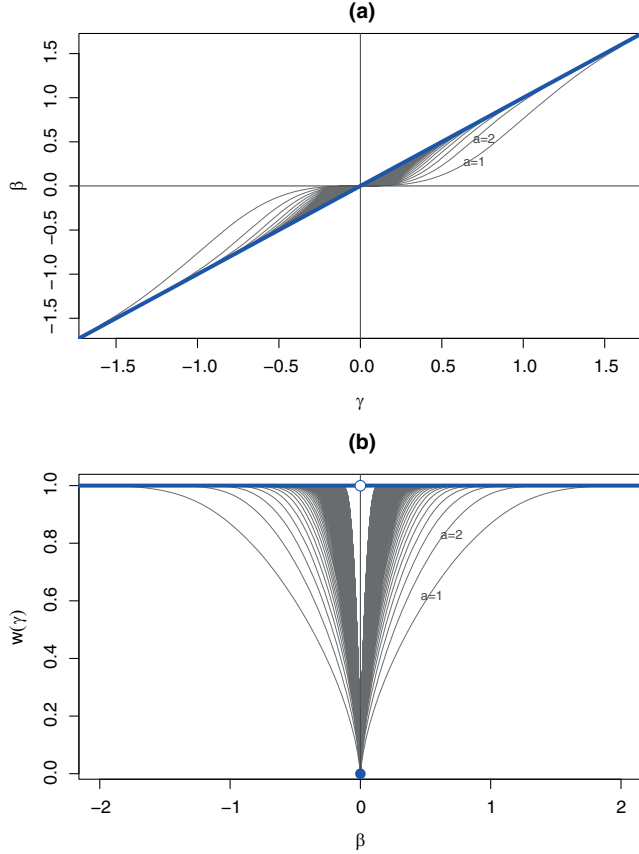


Figure 2. Illustration of the reparameterization step: (a) plot of β vs. γ and (b) plot of $w(\gamma)$ as a function of β , where $w(\gamma) = \tanh(a\gamma^2)$ and $\beta = \gamma w(\gamma)$ for $a = 1, 2, \dots, 200$.

of $\min_{\beta} Q_n(\beta)$ with the objective function

$$Q_n(\beta) = -\frac{2}{n} \cdot l(\beta) + \frac{\ln(n_0)}{n} \cdot \sum_{j=1}^p \rho_n(\beta_j), \quad (6)$$

where the penalty function $\rho_n(\beta_j)$ is defined through the reparametrization $\beta_j = \gamma_j w(\gamma_j)$ and $\rho_n(\beta_j) = w(\gamma_j) = \tanh(a_n \gamma_j^2)$. Furthermore, we assume that $a_n = O_p(n)$.

Let β_0 denote the true sparse parameter vector and partition it as $\beta_0 = (\beta_{0(1)}^T, \beta_{0(2)}^T)^T$, where $\beta_{0(1)} \in \mathbb{R}^q$ consists of all q nonzero components and $\beta_{0(2)}$ consists of all the $(p - q)$ zero components. Let $Y_i(t) = I\{T_i \geq t\}$ be the at-risk process. Define

$$S^{(k)}(\beta, t) = \frac{1}{n} \sum_{i=1}^n Y_i(t) \exp\{\beta^T \mathbf{z}_i(t)\} \mathbf{z}_i^{\otimes k}, \quad (7)$$

for $k = 0, 1$, and 2 , where the outer product notation \otimes is operated as follows: $\mathbf{a}^{\otimes 0} = 1$, $\mathbf{a}^{\otimes 1} = \mathbf{a}$, and $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^T$ for any vector \mathbf{a} . Let $s^{(k)}(\beta, t) = E[Y(t) \exp\{\mathbf{z}(t)^T \beta\} \mathbf{z}(t)^{\otimes k}]$ be the expected value of $S^{(k)}(\beta, t)$. Then, the expected Fisher

information matrix associated with the true model is

$$\mathbf{I}(\beta_0) = \int_0^1 \left[s^{(2)}(\beta_0, t) - \frac{\{s^{(1)}(\beta_0, t)\}^{\otimes 2}}{s^{(0)}(\beta_0, t)} \right] h_0(t) dt.$$

The following theorem shows that, under regularity conditions, there exists a local minimizer $\tilde{\beta}$ of $Q_n(\beta)$ that is \sqrt{n} -consistent and this \sqrt{n} -consistent $\tilde{\beta}$ enjoys the “oracle” property.

THEOREM 1. Assume that $\{(T'_i, C'_i, \mathbf{z}_i) : i = 1, \dots, n\}$ are n i.i.d. copies of (T', C', \mathbf{z}) , $T'_i \perp\!\!\!\perp C'_i \mid \mathbf{z}_i$ for each i , and $n_0 = O_p(n)$. Under the regularity conditions (A)–(D) in Anderson and Gill (1982) or Fan and Li (2002), we have

- (i). (\sqrt{n} -Consistency) there exists a local minimizer $\tilde{\beta}$ of $Q_n(\beta)$ such that $\|\tilde{\beta} - \beta_0\| = O_p(n^{-1/2})$.
- (ii). (Sparsity and Asymptotic Normality) Partition the \sqrt{n} -consistent local estimator in (i) as $\tilde{\beta} = (\tilde{\beta}_{(1)}^T, \tilde{\beta}_{(2)}^T)^T$ in a similar manner to β_0 . With probability tending to 1, $\tilde{\beta}$ must satisfy that $\tilde{\beta}_{(2)} = \mathbf{0}$ and

$$\sqrt{n}(\tilde{\beta}_{(1)} - \beta_{0(1)}) \rightarrow N\{\mathbf{0}, \mathbf{I}_{11}^{-1}(\beta_0)\},$$

as $n \rightarrow \infty$, where $\mathbf{I}_{11}(\beta_0)$ is the leading $q \times q$ submatrix of $\mathbf{I}(\beta_0)$.

Theorem 1 is analogous to Theorems 1 and 2 in Zhang and Lu (2007). Its proof, deferred to the Supplementary Materials, follows Fan and Li (2002) in principle. Nevertheless, since there is no further flexibility offered by adjusting the tuning parameter as in SCAD or ALASSO, properties of the hyperbolic tangent penalty also play a critical role in the proof.

Theorem 1 offers a way of computing the standard errors (SE) for nonzero components $\tilde{\beta}_{(1)}$ in $\tilde{\beta}$. Note that \mathbf{I}_{11} , the leading $q \times q$ submatrix of \mathbf{I} , is exactly the same as the Fisher information matrix associated with the reduced model obtained by eliminating terms associated with zero components $\tilde{\beta}_{(2)}$. An alternative sandwich SE formula for $\tilde{\beta}_{(1)}$ is also available following arguments similar to Fan and Li (2002), for which we shall not pursue further. However, the SE formulas in both approaches are only available for nonzero MIC estimates. Thus, these practices belong to postselection inference and should be used with caution.

3. Inference on β via γ

Postselection inference is inherent for the best subset selection and regularization due to their two-step estimation process. In MIC, variable selection and parameter estimation are completed in one single optimization step. This offers us a unique opportunity to circumvent this fundamental problem. We achieve this with the aid of reparameterization.

The transformation $\beta = \gamma w(\gamma)$ facilitates an important convenience: inference on β can be made via γ . This is because the mapping between β and γ is a bijection and $\beta = \mathbf{0}$ if and only if $\gamma = 0$. Therefore, testing $H_0 : \beta_j = 0$ is equivalent to

testing $H_0 : \gamma_j = 0$. For a zero estimate $\tilde{\beta}_j$, we cannot compute its standard error. But the objective function remains smooth in $\boldsymbol{\gamma}$. The statistical properties of $\tilde{\boldsymbol{\gamma}}$ are readily available following standard M-estimation arguments. In particular, its asymptotic normality is given in the following theorem.

THEOREM 2. *Let $\boldsymbol{\gamma}_0$ be the reparameterized parameter vector associated with $\boldsymbol{\beta}_0$. Under the regularity conditions (A)–(D) in Anderson and Gill (1982), we have*

$$\sqrt{n}[\mathbf{D}(\boldsymbol{\gamma}_0)(\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) + \mathbf{b}_n] \xrightarrow{d} \mathbf{N}\{\mathbf{0}, \mathbf{I}^{-1}(\boldsymbol{\gamma}_0)\}. \quad (8)$$

where

$$\mathbf{D}(\boldsymbol{\gamma}_0) = \text{diag}(w_j + \gamma_j \dot{w}_j) \Big|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}_0} = \text{diag}(D_{jj}), \quad (9)$$

and the asymptotic bias

$$\mathbf{b}_n = \left\{ -\ddot{L}(\boldsymbol{\beta}_0) \right\}^{-1} \frac{\ln(n_0)}{2} \left(\frac{\dot{w}_j}{w_j + \gamma_j \dot{w}_j} \right)_{j=1}^p = (b_{nj}), \quad (10)$$

satisfy (i) $\lim_{n \rightarrow \infty} D_{jj} = I\{\gamma_{0j} \neq 0\}$ and (ii) $\mathbf{b}_n = o_p(1)$.

The proof of Theorem 2 is given in the Supplementary Materials. Several comments are in order. Accordingly, an asymptotic $(1 - \alpha) \times 100\%$ confidence interval for $D_{jj} \gamma_{0j}$ can be simply given as

$$(\tilde{D}_{jj} \tilde{\gamma}_j + \tilde{b}_{nj}) \pm z_{1-\alpha/2} \sqrt{(\mathbf{I}^{-1}(\tilde{\boldsymbol{\gamma}})/n)_{jj}}, \quad (11)$$

where \tilde{D}_{jj} is an estimate of D_{jj} by replacing γ_{0j} with $\tilde{\gamma}_j$ and similarly for \tilde{b}_{nj} and $\mathbf{I}^{-1}(\tilde{\boldsymbol{\gamma}})$. Empirically, we replace the expected Fisher information matrix $\mathbf{I}(\tilde{\boldsymbol{\gamma}})$ with the observed Fisher information matrix $\mathbf{I}_n(\tilde{\boldsymbol{\gamma}})$ given by

$$\mathbf{I}_n(\tilde{\boldsymbol{\gamma}}) = -\nabla^2 l(\tilde{\boldsymbol{\gamma}}) = \sum_{i=1}^n \delta_i \left\{ \frac{S^{(2)}(\tilde{\boldsymbol{\gamma}}; T_i)}{S^{(0)}(\tilde{\boldsymbol{\gamma}}; T_i)} - \left(\frac{S^{(1)}(\tilde{\boldsymbol{\gamma}}; T_i)}{S^{(0)}(\tilde{\boldsymbol{\gamma}}; T_i)} \right)^{\otimes 2} \right\},$$

where functions $S^{(k)}(\tilde{\boldsymbol{\gamma}}; T_i)$ are defined earlier in (7). Note that it is computationally advantageous to use $\mathbf{I}_n^{-1}(\tilde{\boldsymbol{\gamma}})$ rather than $\mathbf{I}_n^{-1}(\tilde{\boldsymbol{\beta}})$, although both $\tilde{\boldsymbol{\gamma}}$ and $\tilde{\boldsymbol{\beta}}$ are consistent for $\boldsymbol{\beta}_0$. Working with $\tilde{\boldsymbol{\beta}}$ entails handling very small or large numbers numerically.

Further simplification of (11) is available by ignoring both \tilde{D}_{jj} and \tilde{b}_{nj} . This is because $D_{jj} \geq 0$ is bounded and equals 0 only when $\gamma_{0j} = 0$. Asymptotically, $\lim_{n \rightarrow \infty} D_{jj} = I\{\gamma_{0j} \neq 0\}$. Moreover, the bias term b_{nj} is $o_p(1)$ with exponential convergence for estimates of the nonzero components in $\boldsymbol{\gamma}_0$ and $O_p\{\ln(n_0)/\sqrt{n}\}$ for estimates of its zero components. Thus, an asymptotic $(1 - \alpha) \times 100\%$ confidence interval for γ_{0j} can be simply given as

$$\tilde{\gamma}_j \pm z_{1-\alpha/2} \sqrt{(\mathbf{I}_n^{-1}(\tilde{\boldsymbol{\gamma}})/n)_{jj}}. \quad (12)$$

Significance testing on $\boldsymbol{\gamma}_0$ can be done in a similar manner.

4. Numerical Studies

In this section, we first discuss numerical and optimization issues in implementing MIC, then present simulation studies that are designed to assess the performance of MIC and compare it to other available methods. We also explore the standard error formula for nonzero components in $\boldsymbol{\beta}$ and inference on $\boldsymbol{\beta}$ via $\boldsymbol{\gamma}$. Finally, a real data example illustration is provided via analysis of the PBC data. Additional numerical results are presented in the Supplementary Materials.

4.1. Implementation Issues

The asymptotic results in Section 2.3 entail $a_n = O_p(n)$. In all the reported numerical results throughout the article, we have set $a_n = n_0$, i.e., number of observed deaths in the data, because $n_0/n \xrightarrow{P} \Pr\{C' \geq T'\}$ by WLLN. In summary, MIC have the following simple form

$$\min_{\boldsymbol{\gamma}} -2l(\boldsymbol{\beta}) + \ln(n_0) \sum_{j=1}^p w_j, \quad (13)$$

where $w_j = \tanh(n_0 \gamma_j^2)$ and $\boldsymbol{\beta} = (\beta_j) = (\gamma_j w_j)$. We would like to emphasize that a is not a tuning parameter as important as the penalty parameter λ . In fact, the MIC estimate stays rather invariant with the choice of a , as demonstrated with additional numerical results presented in the Supplementary Materials. Comparatively, a small change in λ can dramatically change the final estimate and hence fine tuning is necessary in other regularization methods.

The MIC formulation of (13) leads to a smooth programming problem. Nevertheless, the unit dent function is nonconvex in nature. This implies that (13) may have multiple local optima. In our implementation, we tried to make efficient use of readily available optimization routines. We have found that simulated annealing (SA; Belisle, 1992) followed by a BFGS quasi-Newton algorithm (see, e.g., Nocedal and Wright, 1999), both implemented in R (R Development Core Team, 2015) function `optim()`, is quite efficient and effective in computing MIC estimators. Simulated annealing is a global optimization technique that helps seek the global optimum. Succeeding SA with the BFGS method makes sure that the final estimate converges to a critical point.

4.2. Simulation Results

For the convenience of comparison, we have simulated data from the same models as in Zhang and Lu (2007). A total of $p = 9$ covariates $(Z_1, \dots, Z_9)^T$ are generated from a multivariate normal distribution $\text{MVN}_p(\mathbf{0}, \boldsymbol{\Sigma})$, where the covariance matrix $\boldsymbol{\Sigma}$ is given by

$$\boldsymbol{\Sigma} = (\Sigma_{jj'}) \quad \text{with element} \quad \Sigma_{jj'} = 0.5^{|j-j'|} \quad \text{for } j, j' = 1, \dots, 9. \quad (14)$$

Two models, A and B, were considered with true regression coefficients

$$\text{Model A: } \boldsymbol{\beta} = (-0.7, -0.7, 0, 0, 0, -0.7, 0, 0, 0)^T$$

$$\text{Model B: } \boldsymbol{\beta} = (-0.4, -0.3, 0, 0, 0, -0.2, 0, 0, 0)^T,$$

corresponding to larger and smaller effects, respectively. Two censoring rates, 25 and 40%, and three sample sizes $n = 100$,

Table 1

Comparison of selection methods. Data are generated from Models A and B with 100 simulation runs. Three criteria are used: the mean squared error (MSE); the average model size (Size); and the percentage of correct selections (Correct%).

<i>n</i>	Method	Model A						Model B					
		Censoring rate = 25%			Censoring rate = 40%			Censoring rate = 25%			Censoring rate = 40%		
		MSE	Size	Correct%	MSE	Size	Correct%	MSE	Size	Correct%	MSE	Size	Correct%
100	Oracle	0.0672	3.00	100	0.0947	3.00	100	0.0542	3.00	100	0.0832	3.00	100
	Full	0.2247	9.00	0	0.2948	9.00	0	0.1830	9.00	0	0.2389	9.00	0
	Stepwise	0.1198	3.31	72	0.1690	3.32	69	0.1440	2.08	9	0.1851	1.95	4
	MIC	0.1108	3.43	72	0.1543	3.35	69	0.1469	2.06	8	0.1835	1.95	8
	LASSO	0.1456	5.42	10	0.1989	5.51	8	0.1000	4.20	6	0.1424	3.99	3
	ALASSO	0.1178	4.25	35	0.1632	3.97	45	0.1105	3.47	8	0.1572	3.42	4
	SCAD	0.1427	3.44	63	0.1644	3.28	58	0.1233	3.20	11	0.1455	2.70	15
200	Oracle	0.0341	3.00	100	0.0459	3.00	100	0.0239	3.00	100	0.0294	3.00	100
	Full	0.0863	9.00	0	0.1146	9.00	0	0.0744	9.00	0	0.0980	9.00	0
	Stepwise	0.0412	3.18	84	0.0582	3.22	81	0.0644	2.49	32	0.0886	2.42	21
	MIC	0.0428	3.21	80	0.0741	3.49	67	0.0662	2.75	32	0.0812	2.60	24
	LASSO	0.0627	5.45	8	0.1011	5.76	6	0.0529	5.11	7	0.0671	4.93	10
	ALASSO	0.0462	3.97	49	0.0783	4.19	43	0.0554	4.14	15	0.0690	3.77	16
	SCAD	0.0589	4.10	53	0.0626	3.81	62	0.0637	4.16	15	0.0814	3.75	13
300	Oracle	0.0208	3.00	100	0.0267	3.00	100	0.0153	3.00	100	0.0208	3.00	100
	Full	0.0562	9.00	0	0.0666	9.00	0	0.0469	9.00	0	0.0644	9.00	0
	Stepwise	0.0267	3.17	85	0.0331	3.16	86	0.0355	2.79	55	0.0529	2.70	39
	MIC	0.0279	3.26	78	0.0343	3.22	81	0.0316	2.93	61	0.0495	2.67	39
	LASSO	0.0471	5.60	6	0.0635	5.68	7	0.0288	5.18	9	0.0419	5.20	6
	ALASSO	0.0298	3.69	57	0.0445	4.03	51	0.0305	4.27	27	0.0428	4.05	22
	SCAD	0.0325	3.78	61	0.0402	3.88	59	0.0333	3.87	36	0.0479	3.89	22

200, and 300 are experimented. For each simulated data set, seven methods were applied: the oracle model, the full model, the best subset selection; MIC; LASSO with minimum GCV selection of λ ; ALASSO; and SCAD with BIC selection. All the computations were conducted in R (R Development Core Team, 2015). We documented how each method was implemented in the Supplementary Materials.

For performance measures, we reported the mean weighted squared error (MSE) $(\hat{\mathbf{y}} - \mathbf{y})^T \boldsymbol{\Sigma} (\hat{\mathbf{y}} - \mathbf{y})$ with $\boldsymbol{\Sigma}$ given by (14), the averaged model size (i.e., number of nonzero parameter estimates), and percentage of correct selection. Table 1 presents the results based on 100 simulation runs. It can be seen that MIC performs similarly to BSS. However, BSS (even backward deletion) becomes infeasible for moderately large p . More elaboration on this point will be made in the comparison of computing time (see Section B.1 in the Supplementary Materials). Compared to the regularization methods, MIC performs better in terms of all measures in Model A (the case with stronger signals). With weaker signals (Model B), all methods perform poorly when $n = 100$. As sample size increases, their performances all improve. MIC compares favorably to others in terms of the correct selection rate, but less favorably to LASSO or ALASSO in terms of MSE. This can be explained by the fact that MIC is aimed to achieve minimum BIC via approximation and BIC works best with relatively large samples and stronger signals (see, e.g., McQuarrie and Tsai, 1998).

To investigate the postselection SE formula for nonzero estimates, we compare the actual standard deviation (ASD) of estimated $\hat{\beta}_j$ with the mean SE estimates over 500 simulation.

The results are presented in Table 2, together with the coverage probability. It can be seen that the mean SE values are close to the ASD values in most scenario, except in the case of weak signal (Model B) with small sample size ($n = 100$), where the asymptotic SE is smaller than the ASD to a substantial amount. This is the scenario where MIC does poorly in selecting variables due to the use of BIC. The SE formula performs reasonably well in all scenarios in terms of empirical coverage probabilities, which are all close to the nominal confidence level 95%.

To assess the \mathbf{y} -based inference procedure as proposed in Section 3, we recorded the 95% confidence intervals for each individual γ_j and the p-values associated with the Wald test of $H_0 : \gamma_j = 0$. Since MIC is really fast, we have increased the number of simulation runs to 500. Table 3 presents the coverage probability (CP) of the 95% confidence interval, as well as the empirical power (EP) for testing on nonzero coefficients at the significance level $\alpha = 0.05$. It can be seen that the coverage probabilities of the confidence intervals are around the nominal level of 95% in all scenarios. The proposed significance testing procedure also performs reliably in terms of empirical powers, although its performance deteriorates with smaller sample sizes and weak signals as expected.

4.3. Data Example: A Revisit to PBC Data

To a real data example illustration, we pay a revisit to the primary biliary cirrhosis (PBC) data set that is well-known in the survival analysis literature. Another illustration is provided in the Supplementary Materials (Section B.3) where we applied the proposed methods to a gene expression data set.

Table 2

Standard errors for nonzero MIC estimates. The actual sample standard deviation (ASD) of each $\tilde{\beta}_j$, the mean of the asymptotic standard errors (SE-Mean), and the coverage probability (CP) of the 95% confidence intervals are obtained from 500 simulation runs.

			Censoring 25%			Censoring 40%		
	n		ASD	SE-mean	CP	ASD	SE-mean	CP
Model A	100	$\tilde{\beta}_1$	0.174	0.164	93.19%	0.214	0.183	91.55%
		$\tilde{\beta}_2$	0.195	0.168	92.15%	0.224	0.190	92.94%
		$\tilde{\beta}_6$	0.185	0.156	92.38%	0.223	0.177	92.14%
	300	$\tilde{\beta}_1$	0.091	0.089	94.60%	0.102	0.099	94.20%
		$\tilde{\beta}_2$	0.094	0.091	95.20%	0.113	0.102	92.40%
		$\tilde{\beta}_6$	0.095	0.084	92.60%	0.114	0.095	91.20%
Model B	100	$\tilde{\beta}_1$	0.187	0.146	91.79%	0.228	0.163	91.18%
		$\tilde{\beta}_2$	0.214	0.150	91.30%	0.232	0.167	90.34%
		$\tilde{\beta}_6$	0.202	0.139	84.76%	0.209	0.157	88.57%
	300	$\tilde{\beta}_1$	0.086	0.083	94.80%	0.096	0.092	95.00%
		$\tilde{\beta}_2$	0.098	0.083	93.48%	0.109	0.094	96.07%
		$\tilde{\beta}_6$	0.099	0.075	94.17%	0.113	0.084	93.29%

A description of the PBC study, which has been omitted here, can be found in Dickson et al. (1989). This data set has been analyzed by many authors including both Tibshirani (1997) and Zhang and Lu (2007). Note that the results presented in Tibshirani (1997; Table I on p.390) are based on the standardized predictors while the estimated coefficients in Zhang and Lu (2007; Table 5 on p.699) have been transformed back to the original scales. The results from MIC and several other methods, as presented in Table 4, were based on standardized data.

MIC selects eight variables, which are the same as those selected by the stepwise selection and ALASSO. The SCAD model has eight variables too, yet with slightly different choices. The LASSO model is much larger, having 11 selected variables. The final MIC model fit is nearly identical to the one resulted from the stepwise selection, indicating again that MIC mimics the best subset selection method well. The individual parameter testings based on reparameterized $\boldsymbol{\gamma}$ in MIC, free of postselection inference, also support the selected variables.

Table 3

Inference on $\boldsymbol{\beta}$ via reparameterized $\boldsymbol{\gamma}$ in MIC. The coverage probability (CP) of the 95% confidence interval for each individual parameter and the empirical power (EP) at level $\alpha = 0.05$ are based on 500 simulation runs.

		Model A				Model B			
Censoring		$n = 100$		$n = 300$		$n = 100$		$n = 300$	
rate		CP	EP	CP	EP	CP	EP	CP	EP
25%	γ_1	93.6%	99.8%	95.2%	100.0%	94.2%	78.6%	95.6%	100.0%
	γ_2	96.0%	98.8%	95.4%	100.0%	95.6%	41.6%	97.4%	90.8%
	γ_3	95.4%		96.8%		93.4%		94.0%	
	γ_4	96.4%		97.0%		96.4%		96.2%	
	γ_5	96.2%		98.0%		95.4%		95.4%	
	γ_6	96.4%	99.0%	96.4%	100.0%	97.6%	18.2%	97.2%	62.4%
	γ_7	94.6%		96.2%		96.4%		93.2%	
	γ_8	95.2%		97.0%		95.8%		95.0%	
	γ_9	91.6%		97.0%		95.0%		95.0%	
40%	γ_1	94.8%	95.8%	92.8%	100.0%	93.6%	73.8%	95.2%	99.0%
	γ_2	94.0%	95.8%	96.6%	100.0%	95.2%	40.0%	97.8%	87.0%
	γ_3	95.4%		96.8%		96.0%		96.8%	
	γ_4	95.8%		96.6%		93.2%		96.8%	
	γ_5	96.2%		95.8%		94.4%		95.2%	
	γ_6	96.4%	95.6%	96.6%	100.0%	96.8%	16.8%	99.2%	51.4%
	γ_7	96.8%		95.4%		95.2%		93.2%	
	γ_8	97.0%		97.0%		96.8%		97.0%	
	γ_9	96.8%		96.6%		95.0%		96.2%	

Table 4

Analysis of PBC data. The p -value in MIC is computed via the reparameterized $\tilde{\gamma}$ as discussed in Section 3.

	Full model		MIC			Stepwise		LASSO	ALASSO	SCAD
	$\hat{\gamma}$	SE	$\tilde{\beta}$	SE	p-Value*	$\hat{\gamma}$	SE			
trt	-0.062	0.108	0.000		1.0000					
age	0.304	0.123	0.331	0.107	0.0067	0.330	0.107	✓	✓	✓
sex	-0.120	0.103	0.000		1.0000			✓		
ascites	0.022	0.098	0.000		1.0000			✓		
hepato	0.013	0.126	0.000		1.0000					
spiders	0.046	0.111	0.000		1.0000			✓		
edema	0.273	0.107	0.222	0.094	0.0368	0.222	0.094	✓	✓	✓
bili	0.368	0.117	0.391	0.089	0.0006	0.392	0.089	✓	✓	✓
chol	0.116	0.104	0.000		1.0000					✓
albumin	-0.300	0.125	-0.290	0.110	0.0201	-0.291	0.110	✓	✓	✓
copper	0.220	0.103	0.252	0.087	0.0165	0.252	0.087	✓	✓	✓
alk.phos	0.002	0.084	0.000		1.0000					
ast	0.231	0.111	0.248	0.103	0.0276	0.248	0.102	✓	✓	✓
trig	-0.064	0.087	0.000		1.0000					
platelet	0.084	0.110	0.000		1.0000					
protime	0.234	0.107	0.229	0.102	0.0283	0.229	0.102	✓	✓	
stage	0.388	0.150	0.369	0.124	0.0124	0.370	0.124	✓	✓	✓

5. Discussion

MIC offers a new perspective for conducting sparse estimation by approximating a model selection criterion. Su (2015) first experimented a preliminary version of this method in linear regression for the variable selection purpose only, while the current research comprehensively examines sparse estimation within the context of Cox PH models. The main advantages of MIC are summarized as follows. First of all, MIC is free of tuning owing to its special formulation. As a result, MIC is computationally more efficient than many other competitors. MIC only entails the same level of computational complexity as what one would encounter in computing the MPLE. Secondly, BIC is optimal not only for its selection consistency but also because it is derived as an approximation to the posterior distribution of candidate models. The latter property renders the penalty parameter in BIC, i.e., $\lambda_0 = \ln(n_0)$, unique in some sense. This is why BIC is often used as an ultimate yardstick in many variable selection procedures. MIC mimics BSS but extends well to large p scenarios. MIC is also advantageous to regularization methods as it seeks to optimize an approximated BIC without reducing the search space. Even if the fitting algorithm does not guarantee to identify the true global optimum, the final MIC result should correspond to a competitive model with a relatively small BIC. Thirdly, the reparameterization step not only yields computational advantages but also facilitates a leeway to circumvent the fundamental postselection inference problem.

While all our discussions in this article have been restricted to fixed finite dimensions, the general approximation idea of MIC may be extended to scenarios with diverging number of parameters (i.e., $p \rightarrow \infty$ yet $p/n \rightarrow 0$) or ultra-high dimensions with $p \gg n$, where various extended, modified, or generalized information criteria are available as pioneered by Chen and Chen (2008) and others. As a part of our ongoing research efforts, we are investigating how to obtain a modified information criteria for high-dimensional Cox PH models so

that MIC can be readily applied. For future research, MIC can also be possibly applied to other survival models (e.g., accelerated failure time models and frailty models) and various sparsity structures (e.g., grouped or fused LASSO).

6. Supplementary Materials

Proofs and additional numerical results referenced in Sections 2, 3, and 5, as well as the R source codes for computation, are available with this article at the *Biometrics* website on Wiley Online Library.

ACKNOWLEDGEMENTS

The authors thank the Editor, Professor Yi-Hau Chen, for his helpful and constructive comments that have greatly improved an earlier draft. XS was partially supported by NIMHD grant 2G12MD007592 from NIH.

REFERENCES

- Akaike, H. (1974). A new look at model identification. *IEEE Transactions on Automatic Control* **19**, 716–723.
- Anderson, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *The Annals of Statistics* **10**, 1100–1120.
- Belisle, C. J. P. (1992). Convergence theorems for a class of simulated annealing algorithms. *Journal of Applied Probability* **29**, 885–895.
- Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013). Valid post-selection inference. *The Annals of Statistics* **41**, 802–837.
- Blumensath, T. and Davies, M. E. (2009). Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis* **27**, 265–274.
- Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics* **37**, 373–384.

- Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95**, 759–771.
- Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34**, 187–220.
- Cox, D. (1975). Partial likelihood. *Biometrika* **62**, 269–276.
- Dickson, E., Grambsch, P., Fleming, T., Fisher, L., and Langworthy, A. (1989). Prognosis in primary biliary cirrhosis: model for decision making. *Hepatology* **10**, 1–7.
- Fan, J. and Li, R. (2001). Variable selection via nonconcave penalised likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.
- Fan, J. and Li, R. (2002). Variable selection for Cox’s proportional hazards model and frailty model. *Annals of Statistics* **30**, 74–99.
- Fu, W. (1998). Penalized regressions: The Bridge versus the Lasso. *Journal of Computational and Graphical Statistics* **7**, 397–416.
- Furnival, G. M. and Wilson, R. W. (1974). Regression by leaps and bounds. *Technometrics* **16**, 499–511.
- Leeb, H. and Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory* **21**, 21–59.
- Lockhart, R., Taylor, J., Tibshirani, R., and Tibshirani, R. (2014). A significance test for the LASSO. *The Annals of Statistics* **42**, 413–468.
- McQuarrie, A. D. R. and Tsai, C.-L. (1998). *Regression and Time Series Model Selection*. Singapore: World Scientific.
- Nocedal, J. and Wright, S. J. (1999). *Numerical Optimization*. New York, NY: Springer.
- Osborne, M., Presnell, B., and Turlach, B. (2000). On the lasso and its dual. *Journal of Computational and Graphical Statistics* **9**, 319–337.
- R Development Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, <http://www.R-project.org>.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**, 461–464.
- Su, X. (2015). Variable selection via subtle uprooting. *Journal of Computational and Graphical Statistics* **24**, 1092–1113.
- Tibshirani, R. (1997). The LASSO method for variable selection in the Cox Model. *Statistics in Medicine* **16**, 385–395.
- Vollinsky, C. T. and Raftery, A. E. (2000). Bayesian information criterion for censored survival models. *Biometrics* **56**, 256–262.
- Zhang, H. and Lu, W. (2007). Adaptive Lasso for Cox’s proportional hazards model. *Biometrika* **94**, 691–703.

Received July 2015. Revised December 2015.

Accepted December 2015.